

# Models for Decision Support in Healthcare

Kunal Malhotra

College of Computing, Georgia Institute of Technology, Atlanta, Georgia, USA  
kmalhotra7@gatech.edu

**Abstract.** One of the many challenges in the field of medicine is to make the best decisions about optimal treatment plans for patients. Medical practitioners often have differing opinions about the best treatment among multiple available options. While standard protocols are in place for the first and second lines of treatment for most diseases, a lot of variation exists in the treatment plans subsequently chosen. We propose to extensively study recommended treatment guidelines and plans for selected rare and chronic diseases. As representative diseases we study Glioblastoma Multiforme (brain cancer) which is classified as a rare disease, and Diabetes Mellitus, which is a nationally and globally widespread chronic disease. A graph model is designed to capture the data pertaining to the treatment options and historical evidence and further analyzed to discover sequential treatment patterns based on different outcome classes based on longevity, complications etc. The notion of ‘Patient Similarity’ would be explored to form cohorts of clinically similar patients. The treatment patterns would be ranked, and highly ranked patterns would be ordered depending on expected outcomes before being assigned to cohorts of patients. A prototype decision support system is planned for recommending treatment options based on a patient’s clinical profile. Evaluation of the models involves using historical data with various evaluation metrics and also by a qualitative assessment by expert physicians.

**Keywords:** Graph Data, Sequential Pattern Mining, Patient Similarity

## 1 Introduction

Evidence-based medicine refers to the explicit and exhaustive use of available medical evidence to improve quality of care provided. It involves integration of individual clinical expertise with the best available external clinical evidence from systematic research [17]. When deciding on the treatments for patients, medical practitioners consult both clinical evidence and use their own clinical expertise and experience to inform their decisions and recommendations to patients. Variability in such recommendations among multiple practitioners could directly affect a patient’s wellbeing and recovery. A key challenge in the field of medicine, faced by a physician, is determining the optimal treatment for a given patient. While protocols exist for first and second line treatments for the vast majority of diseases, tailoring treatments to an individual patient presents a huge diagnostic challenge. With the ever-expanding treatment options, there

is an increasing need for a system to identify treatment patterns and new drugs that would cure the patient in the most efficient manner. Treatment is also a dynamic process, and must evolve as new information becomes available such as response to a particular treatment approach which may be positive or adverse during the course of treatment.

In acute conditions, the goal is to find a “cure” or treatment to reverse or at least arrest the progression of disease typically manifested in a variety of aggressive cancers. On the other hand, in chronic diseases like diabetes or renal disease, the goal is to manage the disease state and prevent further deterioration or progression to more serious conditions. Part of chronic disease management is to identify and prescribe measures for improving the quality of life. Toward this end, we have selected two specific diseases for analysis, namely Glioblastoma Multiforme (GBM), and Diabetes Mellitus. GBM is the most lethal type of brain cancer and is biologically the most aggressive subtype of malignant gliomas. The current standard of care for GBM patients involves surgical resection followed by radiation and chemotherapy with an oral alkylating agent Temodar [14]. Most patients with GBM survive less than a year after diagnosis. This extreme mortality rate, where none have a long-term survival, has drawn significant attention to improving treatment for these tumors. GBMs are not cured by surgery due to the complex nature of the tumor and variable location of the tumor cells in the brain, resulting in the inability to completely resect this tumor [8]. Chronic diseases are the worlds leading cause of death. Diabetes is a complex disease, often found to co-occur with other chronic conditions such as hypertension and depression, which not only complicates diabetes management and but also increases the risk of developing diabetes by 60% [3]. Majority of patients with diabetes in the U.S. do not achieve the recommended guidelines. [18]. Disparities in treatment are present as well: significantly fewer Black and Latino adults, low-and middle-income adults, and uninsured adults received recommended diabetes care compared with white, high-income, and insured adults between 2002 and 2007 [2].

With the enforcement of electronic medical records(EMRs), a vast amount of healthcare data is being captured. Using various techniques in data mining we can improve decision-making. The results of healthcare data analysis can influence cost, revenue and quality of care [10]. Decision trees have been used extensively based on extensive medical data and clinical evidence to develop decision support systems and stored in a relational database [21]. For this dissertation, we aim at using techniques of information processing and data mining to make use of historical data about patients, guidelines of treatments given to them, health outcomes, complications, etc to develop a set of algorithms and implement them in a decision support tool.

## 2 Problem Definition and Goals of Research

We propose to develop a framework to evaluate efficiency of treatments in terms of outcomes using clinical patient data. Outcomes of interest include likelihood

of survival, longevity, probability of hospital admissions within one year, probability of hospital readmission within 30 days, etc. Short and long term economic outcomes are equally important. We identify characteristic treatment patterns and associate these patterns with patients having similar outcomes.

So far we have been working with GBM which is a rare disease. Our scope covers chronic diseases like Diabetes Mellitus to investigate the validity and accuracy of prediction in both cases. After preliminary discussions with domain experts regarding different cancers and diabetes we anticipate that our general approach will be equally applicable in multiple acute and chronic clinical situations where preliminary and secondary treatments are well understood, but subsequent treatment may have too many options. A systematic decision support methodology would be best suited for both chronic and rare diseases. We identify clinical and behavioral similarities among patients and create patient cohorts based on the similarities found. A clinically relevant distance measure needs to be developed to perform patient profiling [24]. This ‘patient similarity assessment needs extensive exploration for studying the treatments prescribed to different cohorts and help in treatment comparison, management of patients in groups and prediction. Heuristics would be developed to rank the identified treatment patterns for an individual patient based on the extent of similarity a new patient shares with classes of patients. The model would also be enhanced to incorporate various responses to treatments.

The goals of my research are as follows:-

1. **Understanding the disease:** A comprehensive understanding of how the disease affects people, the preferred first and second line of treatments, variations in standard of care, treatment options other than the standard of care, and the circumstances under which a particular course of treatment is prescribed.
2. **Creating a decision support tool:** Models would be built to assist medical professionals in deciding the most probable treatment patterns for individual patients by learning from historical data. There are three subgoals here: a) Recognition of treatment patterns characteristic of a particular outcome or multiple outcomes; b) Defining a similarity measure for patients based on their clinical characteristics; and c) Developing a ranking algorithm to order the treatment patterns to achieve the best outcomes for individual patients or patient cohorts based on their similarity to classes of patients.
3. **Evaluation of the decision support tool** Evaluation of the models would be done by training and testing our algorithms on datasets provided by our partner institutions in this study. The treatments prescribed by the doctors would be compared with the one recommended and ranked by our tool. Validation of our approach would be done using various measures such as receiver operating characteristic curve, accuracy, precision, and recall.

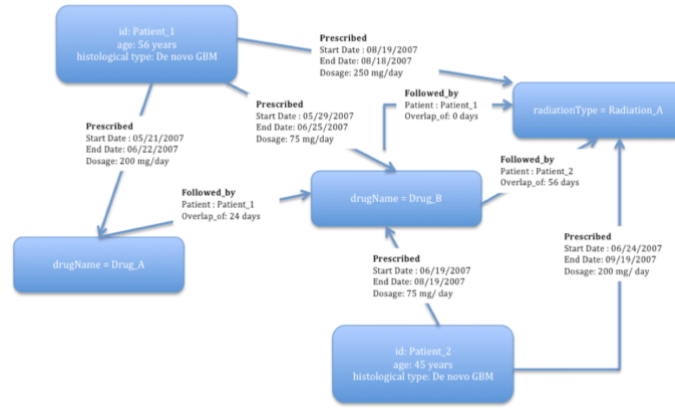
### 3 Methods and Techniques

#### 3.1 Data Representation

A prototype model is being developed for GBM patients using clinical and genomic data from a public portal called ‘The Cancer Genome Atlas Portal’ [12] and cBioPortal [4]. The clinical domain includes demographic information about the patient along with some basic clinical features, e.g. Karnofsky performance score, histological type, survival duration, prior glioma information and most importantly the vital status of the patient (Living / Dead). Studies show that GBM patients can be classified into four subtypes namely Classical, Mesenchymal, Proneural and Neural based on the expression levels of a particular set of genes [7, 23]. For our study we considered these set of genes and used their mRNA expression levels, copy number variation data and methylation status. Additional information includes drugs prescribed along with their dosage, therapy type, radiation type, radiation dosage, and start and end dates for the treatment. We model this data as a graph where nodes are of two types: ‘patient node’ & ‘treatment type node’ and edges are also of two types: ‘prescription edge’ & ‘sequence edge’. A graph offers a much richer picture of a network, and relationships of several types. Since the data model has a path-oriented nature, the majority of path-based graph database operations are highly aligned with the way in which the data is laid out hence increasing the efficiency [16]. Figure 1 shows the current representation of the data as a graph. The figure shows a graph consisting of two patients just for illustrative purposes. In the graph patient nodes have properties such as ‘patient id’, ‘age’, etc. Drugs and radiation prescribed are represented as treatment type nodes with properties such as ‘drug name’ and ‘radiation type’ respectively. The ‘prescription edge’ signifies the prescription of treatment with properties such as ‘start date of prescription’, ‘end date of prescription’, ‘dosage’, etc. The ‘sequence edge’ signifies the sequence in which drugs or radiation were prescribed. E.g., The edge labeled ‘Prescribed’ between the patient node with ‘id = Patient\_1’ and the drug node with ‘drugName = Drug\_A’ signifies that ‘Patient\_1’ was prescribed 200 mg/day of ‘Drug\_A’ on 05/21/2007 till 06/22/2007. The other type of edge labeled ‘Followed\_by’ would always be between two drugs or two types of radiation or between a radiation type and a drug signifying the sequence of the prescription. E.g., the ‘Followed\_by’ edge between source node ‘Drug\_A’ and target node ‘Drug\_B’ with properties ‘patient’ and ‘overlap’ signifies that for ‘Patient\_1’, Drug\_A was followed by Drug\_B and there was an overlap of 24 days. The graph shown in the above figure is based on the data available for GBM patients. The structure of the graph would be enhanced for the Diabetes patients since the chronicity of the disease and its commonly comorbid state with other conditions such as hypertension and clinical depression yields many more parameters and potential complications to consider.

#### 3.2 Approach

Our approach is driven by the outcome of the treatment. If survival is the outcome, the objective is to increase the survival period of the patient as much



**Fig. 1.** Data represented as a graph

as possible. For GBM patients we analyze the treatment data of patients and identify treatment patterns, which are characteristic of survival for a certain range of time (e.g., 6-12 months). The notion of ‘patient similarity’ would be explored to either use existing similarity measures or develop a new metric for the purpose of identifying similar patients which would play a significant role in recommending treatment for a cohort of patients. Significant work has been done at the Healthcare Systems and Analytics Research department of IBM in the area of patient similarity involving physician feedback as an important parameter to group similar patients [22]. Chan et al (2010) [5] have proposed a new patient similarity algorithm named SimSVM, which does a binary classification and outputs the predicted class which is survival greater than 12 months or less than 12 months and degree of similarity or dissimilarity. Their approach only considers a single outcome and a few similarity measures as input. We believe that our approach will be a significant enhancement since we plan to consider multiple outcomes together as we believe that a single outcome based approach could be misleading. For each individual patient we would rank the treatment patterns based on historical experience with similar patients. Heuristics would be developed to do the ranking.

We categorize the patients based on some outcome variable as a range of values: e.g., survival period of ‘less than 6 months’, ‘6-12 months’, ‘12-18 months’, etc. The treatments for all the patients in each period would be represented as a graph. Sequential pattern mining techniques such as GSP (Generalized Sequential Pattern mining) [1] & SPADE (Sequential Pattern Discovery using Equivalence classes) [25] have been applied to come up with a “predominant” treatment pattern for every period. This pattern may consist of a combination of multiple drugs following a sequence and would be characteristic of a particular survival period representing the most commonly used treatment associated with that period. These techniques are motivated by association rule mining techniques such as the Apriori algorithm. Initially a combination of N=2 drugs

following a sequence are accounted for and the ones prescribed to a significant number of patients are considered for further analysis where 'N' is the number of drugs. This is followed by a combination of N+1 drugs and so on and so forth. The algorithm terminates when no more significant combinations can be formed. These patterns are used as features to classify patients based on survival periods.

The goal is to find patterns best suited for a particular profile of patients sharing similar clinical and/or genomic characteristics. For a recently diagnosed patient we would filter out patient profiles which share similarity above a particular threshold. Even though all the candidate patient profiles have clinical similarity with the test patient, their degree of similarity may vary and would be reflected through the distance metric that we propose to use. Due to the difference in the extent of patient similarity, we would assign weights to the treatment patterns prescribed to patients belonging to a particular profile. These weights would be dependent on the following parameters:

1. Distance between the test patient and the candidate patient class: The weight assigned to treatment patterns characteristic of a particular patient class is inversely proportional to the distance from the test patient than another candidate class.
2. Number of candidate patients following a particular type of treatment pattern with respect to a particular profile: The larger the number of patients following a particular treatment pattern, the higher the weight. A particular treatment pattern may have different weights for different patient profiles.
3. Other criteria could include the degree of side effects, the extent of adverse reactions in patients and several other clinical considerations as determined by some consensus among physicians

## 4 Preliminary Results

Preliminary analysis on the GBM data was performed and a classification model to predict the survival period of patients was developed. Our goal in this analysis was to perform a binary classification of the patients into 2 classes based on survival period; namely, a) patients surviving more than 1 year b) patients surviving less than 1 year and report the features, which are predictive of the survival duration. A linear classifier (Logistic Regression) was trained based on the clinical and the genomic features recorded before the treatment was started and the sequential patterns which were extracted based on the treatment administered to the patients within one year of their diagnosis. For this experiment we did not consider the drugs prescribed as standard of treatment to train the classifier. Forward feature selection was used to pick predictive features for the model. The Area under the receiver operating characteristic curve (AUC) [19] and accuracy used for evaluation of the model are reported in Table 1. The predictive features influencing the survival period of GBM patients are shown in Table 2.

**Table 1.** Table 1. Performance of various feature combinations in predicting the survival period of GBM patients

Model	AUC	Accuracy
Clinical Features	0.66 +/- 0.166	68% +/- 9.33%
Treatment Patterns	0.69 +/- 0.059	70% +/- 8.08%
Genomic Features	0.70 +/- 0.10	70% +/- 10.49%
Genomic Features and Treatment Patterns	0.76 +/- 0.078	75.7% +/- 5.72%
Treatment Patterns and Clinical Features	0.77 +/- 0.078	74.7% +/- 6.27%
Genomic Features and Clinical Features	0.78 +/- 0.048	76.38% +/- 4.56%
Genomic Features and Treatment Patterns and Clinical Features	0.80 +/- 0.07	78.6% +/- 8.27%

**Table 2.** Table 1. Predictive Features influencing the survival period of GBM patients to be greater than one year

Predictive Features
<b>Genomic</b>
EFGR copy number variation = -1 (hemizygous deletion)
GABRA1 mRNA expression z-score between 1 and -1
OLIG2 mRNA expression z-score between -2 and -2.5
PDGfra mRNA expression z-score between -1.5 and -2
RELB mRNA expression z-score between -1.5 and -2
SYT1 mRNA expression z-score between 1 and 1.5
TP53 mRNA expression z-score between -1.5 and -2.5
<b>Clinical</b>
Age of patient at pathologic diagnosis
Karnofsky performance score
<b>Treatment Patterns</b>
External Beam Radiation therapy followed by prescription of Temodar followed by Avastin
Prescription of Dexamethasone
Prescription of Temodar followed by CCNU
Prescription of Temodar followed by External Radiation followed by Dexamethasone

## 5 Related Work

Some work has been done in the area of developing models for predicting treatment plans for patients. Research groups have developed models to predict the various drug interventions as well as drugs coupled with lab interventions that would work best for a particular disease. These models do not include some important parameters like symptoms, results of investigations, laboratory test results, etc. The treatment plan predicted by these models is only limited to drugs that may be effective [15]. We consider a very comprehensive definition of a treatment plan and the approach outlined previously would predict, besides important drugs, the dosage of each drug, duration of drug therapy, types of interventions, duration of a particular intervention, etc. tailored to a particular patient profile. Based on the models built by Kim. et al (2004) [9] for chronic heart failure (CHF) treatment, significant factors improving the plasma BNP

levels were discovered . which were validated by large-scale trials. Similar work has been done in the area of heart disease diagnosis reporting fairly good accuracy [20]. Neuvirth et al (2011) [13] present a prototype for a data-driven risk assessment system for Diabetes patients and claim to identify physicians who can deliver optimal care to such patients and also identify patients requiring emergency care services. The first model predicting high risk patients was a binary classification to predict whether a patient attended ‘Urgent Care’ services in a specific period of time. The AUC (Area under the curve) used to validate the classification ranged from 0.57–0.71 which does not seem to be a very significant classification. The second model predicting best physician match, uses features that categorizes the match of the patient and the physician by quantifying the similarity of the patient to the characteristic population of the physician. The AUC for this model is approximately 0.59 and based on this result the authors claim that personalized physician assignment plays a role in the treatments success. The decision support model developed by Chen et al(2012) for Diabetes [6] uses a case based reasoning approach to find patient cases similar to the one queried and is not very robust since the approach used by authors to find similar cases is not very granular and eventually the same line of treatment which is given to these similar cases is recommended for the new patient. In our approach we will take into consideration all the cohorts of patients similar to the test patient and then assign weights to the treatment pattern in each cohort, which we believe would be more accurate than the case based reasoning approach. We extensively tease out the different treatment patterns that are characteristic of a particular outcome and plan to come up with a meaningful measure of patient similarity to build patient profiles based on clinical and possibly behavioral variables, especially for diabetes. We believe our approach to assign weights to different treatment patterns for a particular patient profile is also very comprehensive and unique especially because we will be relying on consensus from domain experts for individual diseases.

## 6 Expected Outcomes

Most of the work mentioned below will feed into the prototype development work. Actual testing will be done for the validation of algorithms with available data. Overall testing of our treatment suggestion system will be done in conjunction with the prototype implementation.

**Models:** A graph structure is being used to capture GBM data but keeping in mind the extensive nature of chronic disease data, some changes might have to be made to the existing graph structure. We are currently exploring various sequential pattern mining techniques and evaluating them given the complex nature of the graphs we have [11]. These models are subjected to a variety of analyses. Our overall approach can be considered as a predictive model for capturing historical patient treatment data and coming out with a ranked scheme for possible treatments that would be targeted to certain desired outcomes.



**Algorithms:** A variety of algorithms will be designed including sequential pattern mining algorithms to analyze treatment data with multiple parameters, clustering algorithms to cluster patients into patient cohorts, algorithms to define a patient clinically in terms of patient similarity or to define the distance of a patient from a patient class. In addition to these, weighting and ranking algorithms for treatments would also be designed.

## 7 Contribution of Dissertation

The dissertation is focussed on developing a decision support model for rare and chronic diseases and finding treatment patterns which can influence a particular outcome or multiple outcomes. In the paper we presented the first step in mining significant sequential treatment patterns and correlating the patterns with survival period of patients. From a clinical perspective, it gives an insight to the medical community about the significance of prescribing a set of drugs in a particular sequence. Our current approach unlike previous approaches takes into account duration of prescription of drugs, overlapping prescriptions, etc. We would also incorporate decision making based response shown by patients to the treatments recommended under different circumstances. We are currently exploring the area of patient similarity to form cohorts of patients sharing similar clinical features. Each cohort of patients may have different treatment patterns which are more predominant than the other which would help us in assigning weights to the treatment patterns in addition to inputs from domain experts. Based on the level of similarity shared by a newly diagnosed patient we would recommend treatment guidelines. Using a variety of available clinical features about patients and their treatments we plan to develop a comprehensive testbed of decision models.

## References

1. Agrawal R, Srikant R. "Mining Sequential Patterns". ICDE 1995: 3-14
2. AHRQ. National Healthcare Disparities Report, 2010. Rockville, MD: Agency for Healthcare Research and Quality. 2011. Contract No.: AHRQ Publication No. 11-0005.
3. Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011
4. Cerami et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery. May 2012 2; 401.
5. Chan L.W.C, Chan T., Cheng L.F, Mak W.S (2010). "Machine Learning of Patient Similarity: A Case Study on Predicting Survival in Cancer Patient after Locoregional Chemotherapy" 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops.

6. Chen J.X , Su S.L, Chang.C.H (2012). "Diabetes Care Decision Support System". 2nd International Conference on Industrial & Information Systems.
7. Hegi et al (2005). "MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma." The New England Journal Of Medicine 352(10): 997-1003.
8. Holland E.C (2000). "Glioblastoma Multiforme: The terminator." PNAS 97(12): 6242-6244.
9. Kim.J et al (2004). "A Novel Data Mining Approach to the Identification of Effective Drugs or Combinations for Targeted EndpointsApplication to Chronic Heart Failure as a New Form of Evidence-based Medicine." Cardiovascular Drugs and Therapy 18(6): 483-489.
10. Kon H.C, Tan G. (2005). "Data mining applications in healthcare." Journal of Healthcare Information Management 19(2): 64-72.
11. Mooney C.H, Roddick J.F. "Sequential Pattern Mining - Approaches and Algorithms". ACM Computing Surveys, Vol.45, No.2, Article 19, Feb 2013. Sequential Pattern Mining Survey paper Carl H Mooney and John F. Roddick.
12. Network, T. R. (2010). The Cancer Genome Atlas Data Portal, National Institute of Health.
13. Neuvirth et. al (2011). "Toward personalized care management of patients at risk: the diabetes case study". KDD: 395-403.
14. Parsons et. al. "An Integrated Genomic Analysis of Human Glioblastoma Multiforme." Science 321(5897): 1807-1812, 2008.
15. Razali A.M et al (2009). "Generating Treatment Plan in Medicine : A Data Mining Approach" American Journal Of Applied Sciences 6(2): 345-351.
16. Robinson.I , Webber.J, Eifrem.E (2013). Graph Databases. California, O'Reilly Media Inc.
17. Sackett D.L, Rosenberg W.M.C, Gray J.A.M, Haynes R.B, Richardson W.S (1996). "Evidence based medicine: what is it and what is isnt." BMJ 312(7023): 71-72.
18. Saaddine J.B, Cadwell B, Gregg E.W, Engelgau M.M, Vinicor F, Imperatore G, Narayan K.M. "Improvements in diabetes processes of care and intermediate outcomes: United States", 1988-2002.[summary for patients in Ann Intern Med. 2006 Apr 4;144(7):112; PMID: 16585656]. Annals of Internal Medicine.2006;144(7):465-74.
19. Seong Ho Park, Goo J.M, Jo C.H (2004). "Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists." Korean Journal of Radiology 5(1): 11-18.
20. Shouman M, Turner T, Stocker R (2012). "Using data mining techniques in heart disease diagnosis and treatment" Japan-Egypt Conference on Electronics, Communications and Computers
21. Skevofilakas M.T et al (2005). "A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines." Conference Proceedings IEEE Engineering in Medicine and Biology, Shanghai China.
22. Sun.J, W. F., Hu.J, Ebadollahi.S, (2012). "Supervised Patient Similarity Measure of Heterogenous Patient Records." SIGKDD Explorations 14(1): 16-24.
23. Verhaak R.G et al (2010). "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1." Cancer Cell 17(1): 98-110.
24. Wang.F, Sun.J, Hu.J, Ebadollahi.S (2011). "iMet: Interactive Metric Learning in Healthcare Applications". SDM: 944-955.
25. Zaki.M.J. "SPADE: An Efficient Algorithm for Mining Frequent Sequences". Machine Learning 42(1/2): 31-60 (2001)